1

На правах рукописи

Глазырин

Глазырин Николай Юрьевич

Алгоритмическое распознавание аккордов в цифровом звуке

Специальность 05.13.18 — математическое моделирование, численные методы и комплексы программ

Автореферат

диссертации на соискание учёной степени

кандидата физико-математических наук

Екатеринбург — 2015

Работа выполнена в ФГАОУ ВПО «Уральский федеральный университет имени первого Президента России Б.Н. Ельцина» на кафедре алгебры и дискретной математики Института математики и компьютерных наук.

Научный руководитель:	Волков Михаил Владимирович доктор физико-математических наук, профессор.
Официальные оппоненты:	Титов Сергей Сергеевич доктор физико-математических наук, профессор, ФГБОУ ВПО «Уральская государственная архитектурно-художественная академия», заведующий кафедрой прикладной математики и технической графики. Харуто Александр Витальевич, кандидат технических наук, доцент, ФГБОУ ВПО «Московская государственная консерватория имени П.И. Чайковского», заведующий кафедрой музыкально-информационных технологий.
Ведущая организация:	ФГАОУ ВПО «Московский физико-технический институт (государственный университет)».

Защита состоится 18 февраля 2015 г. в 13:00 на заседании диссертационного совета Д 212.285.25 на базе ФГАОУ ВПО «Уральский федеральный университет имени первого Президента России Б.Н. Ельцина» по адресу: 620000, г. Екатеринбург, пр. Ленина 51, зал заседаний диссертационных советов, комн. 248.

С диссертацией можно ознакомиться в библиотеке и на сайте ФГАОУ ВПО «Уральский федеральный университет имени первого Президента России Б.Н. Ельцина», http://dissovet.science.urfu.ru/news2/

Автореферат разослан «13» января 2015 года.

Ученый секретарь
диссертационного совета Д 212.285.25
доктор физико-математических наук,
профессор — Пименов В.Г.

Общая характеристика работы

На текущий момент компьютер является основным средством для хранения и обработки музыки и любой информации о музыке, будь то ноты, биография композитора, год выпуска записи или график концертов группы. Необходимость получения разнообразной информации о конкретной цифровой звукозаписи порождает множество задач, связанных с обработкой звука: идентификация композиции, нахождение разных версий одной композиции, определение заданной композиции в потоке звука с радио, поиск похожих композиций, определение мелодии композиции для последующего воспроизведения на музыкальном инструменте и другие. Эта диссертация посвящена задаче автоматического определения последовательности аккордов в цифровом звуке.

Звук можно представлять себе как ограниченную функцию $x(t)$, показывающую давление среды в данной точке в зависимости от времени. Любая музыкальная звукозапись может быть представлена как сумма звуков отдельных музыкальных инструментов. В свою очередь, звук каждого из инструментов разбивается на сумму отдельных звуков, соответствующих нотам, воспроизводимым в разные моменты времени. Каждый из таких отдельных звуков также имеет начало и длительность звучания. Кроме того, каждый музыкальный инструмент имеет тембр. Он определяет набор отдельных гармонических колебаний (частотных компонент), составляющих каждый из звуков, соответствующих нотам.

Задача восстановления по звукозаписи всех отдельных составляющих нот (время начала и окончания звучания, название ноты и номер октавы) называется задачей транскрибирования. В общем виде она не решена. Более простыми являются, например, задачи определения количества и набора инструментов, участвующих в звукозаписи, задача определения всех моментов начала звучания нот, задача определения последовательности аккордов.

Как и в случае транскрибирования, при распознавании последовательности аккордов в звукозаписи, фактически, делается попытка приблизить исходную функцию $x(t)$ суммой отдельных функций, соответствующих аккордам. Поскольку любой аккорд является сочетанием одновременно звучащих нот, которые при этом не требуется разделять по инструментам и отдельным партиям, приближение оказывается менее детальным, чем при транскрибировании. Алгоритм при этом оказывается существенно проще за счёт отказа от излишней детализации.

Задачу определения последовательности аккордов в звукозаписи можно поставить следующим образом: пусть заданы звуковой сигнал $x(t)$, $t \in [t_{start}, t_{end}]$ и множество возможных названий аккордов Y. Необходимо для каждого момента времени $t \in [t_{start}, t_{end}]$ указать аккорд $y \in Y$, звучащий в этот момент.

Система для определения последовательности аккордов в звуке может быть использована при обучении игре на музыкальном инструменте (например, для автоматического аккомпанемента музыканту-любителю), при изучении теории музыки (для иллюстрации закономерностей в музыкальных композициях). Ещё одним потенциально интересным применением является индексирование музыкальных звукозаписей по содержимому с целью дальнейшего поиска. Последовательность аккордов, как правило, является устойчивым инвариантом музыкальной композиции относительно различных аранжировок, поиск которых может быть одним из вариантов использования такой системы.

Актуальность темы. Рассматриваемая задача впервые возникает в 1990-х годах. В первых работах (например, в [2]) она решалась путём предварительного транскрибирования, то есть распознавания отдельных нот и их последующего объединения в аккорды. Т. Фуджишима в [9] предложил способ распознавания аккордов без предварительного определения отдельных нот, который лёг в основу всех последующих методов.

Такие методы обычно организованы следующим образом. Вспомним, что, фактически, музыкальный звук является совокупностью отдельных звуков, соответствующих нотам. Каждый из таких звуков является совокупностью гармонических колебаний, наиболее сильное из которых является основным тоном, а остальные – обертонами. Наличие колебаний с определёнными частотами в определённые промежутки времени может свидетельствовать о звучании аккорда на этих промежутках. Поэтому естественно начинать процесс распознавания с получения частотно-временного представления звукозаписи или её спектрограммы $C_{N\times M}$. Здесь возникает 2 подзадачи: разбиение звукозаписи на фрагменты и вычисление спектра на каждом из них. Далее для каждого вектора из данной последовательности столбцов спектрограммы $\{C_m\}_{m=0}^{M-1}$ необходимо указать аккорд $y \in Y$, соответствующий этому вектору. При этом решается задача классификации. Поэтому здесь главными подзадачами являются выбор метода классификации и нахождение набора преобразований спектрограммы, облегчающего классификацию. Качество решения каждой из отмеченных подзадач вносит весомый вклад в итоговое качество распознавания аккордов, и ни одна из них не может быть пропущена.

Используемые для получения спектра методы возвращают усреднённый результат для всего заданного фрагмента звукозаписи. Поэтому её разбиение на фрагменты необходимо, чтобы определять изменения в спектре с течением времени и реагировать на смену звучащего аккорда. В некоторых работах при разбиении используется информация о ритме музыки (например, в [7]). В отличие от простого последовательного разделения с фиксированным шагом, учёт ритма

позволяет вычислять спектр в тех точках, где музыкальные инструменты звучат наиболее чётко и выраженно, т.е. позволяет улучшить качество распознавания.

Для получения спектра используют дискретное оконное преобразование Фурье ([24], [19], [7]), преобразование постоянного качества (constant-Q преобразование) [4] ([20], [22]) или гребёнки фильтров ([1], [11]).

Преобразования полученной спектрограммы необходимы для её очистки от шумов и обертонов, которые выражаются в виде всплесков энергии на частотах, отличных от частот основных тонов звуков, соответствующих нотам. Результатом преобразований обычно является последовательность векторов меньшей размерности, чем исходные столбцы спектрограммы. Эти векторы называют векторами признаков. За последние 15 лет было предложено множество различных преобразований и типов векторов признаков ([9], [13], [10], [19], [21], [12]).

Для определения звучащего на данном фрагменте аккорда по вектору признаков необходимо классифицировать этот вектор. Наиболее простой способ классификации – определение расстояния от вектора признаков до «идеальных» шаблонных векторов той же размерности, соответствующих аккордам. В качестве результата выбирается аккорд, расстояние до шаблона которого является наименьшим. Фактически, это метод k ближайших соседей для $k = 1$. Такой подход был применён, например, в [13], [22]. Широко используемые в методах распознавания речи *скрытые марковские модели* (СММ) также нашли применение в алгоритмах распознавания аккордов. В отличие от метода ближайшего соседа, они позволяют в явном виде моделировать вероятность перехода между двумя заданными аккордами. СММ использовались во многих работах, например в [24], [1], [12].

Таким образом, в 2000-х годах определение аккордов окончательно выделяется в отдельную задачу. Результаты в этом направлении докладываются на ведущих международных конференциях: International Society for Music Information Retrieval (ISMIR), IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), Sound and Music Computing Conference (SMC). Начиная с 2008 года в рамках ежегодной кампании по оценке методов музыкального информационного поиска MIREX [17] проводятся соревнования среди алгоритмов распознавания аккордов в музыкальном звучании. За это время был достигнут существенный прогресс в качестве распознавания (совокупная продолжительность участков, на которых аккорд распознан правильно, увеличилась с 65% до 80% от длительности композиции). В 2013 году на это соревнование были выставлены более 10 алгоритмов. Таким образом, есть все основания утверждать, что тема диссертации актуальна.

В последние два года практически во всех работах на этапе классификации векторов признаков используются методы на основе СММ. Несмотря на

достаточно высокое качество распознавания аккордов, такие методы имеют свои недостатки, среди которых Де Хаас в [7] выделяет следующие:

- Потребность в большом количестве размеченных данных для обучения. Подготовка таких данных весьма трудоёмка, а сами данные могут сильно разниться для разных стилей музыки, эпох, композиторов.
- Опасность переобучения. Модели с большим количеством параметров наилучшим образом подстраиваются под доступный набор обучающих данных, но непонятно, насколько хорошо они будут подходить для работы с данными не из обучающей выборки.
- Многомерность данных. Она приводит к экспоненциальному увеличению объема данных и времени их обработки, а также к росту необходимого объёма обучающей выборки.
- Недостаточное использование времени. Марковское свойство предполагает зависимость только от предыдущего фрагмента. Но музыкальная композиция зачастую имеет определённую, достаточно протяжённую по времени, структуру, которая не может быть отражена в модели.
- Существуют другие условия, которые также не могут быть выражены в рамках обучаемой модели. Например, это культурный или географический контекст или сложившиеся практики и правила создания музыки.
- Сложность интерпретации модели, оперирующей в большей степени искусственными, математическими, нежели музыкальными конструкциями.

Ещё одним недостатком является то, что упомянутые вероятностные методы хорошо приспособлены для моделирования смены состояния (звучащего аккорда) и хуже – для моделирования продолжительности нахождения в одном состоянии.

Некоторые из этих недостатков могут быть преодолены при помощи методов глубокого обучения, имеющих в основе многослойную нейронную сеть (например, [23]). Каждый из её слоёв предварительно обучается без учителя на неразмеченных данных, после чего требуется лишь небольшое количество размеченных примеров для окончательной подстройки параметров.

Таким образом, разработка метода для распознавания последовательности аккордов, не требующего большого объема данных для обучения, и не предполагающего использования сложной многопараметрической самообучающейся модели, но при этом сопоставимого по качеству результатов с уже существующими методами, является вполне естественной и актуальной. Именно разработка такого метода стала целью для автора данной работы.

Для достижения поставленной цели необходимо было решить следующие задачи:

1. разработать метод для более точного выделения в звуке компонент, соответствующих музыкальным инструментам, с целью улучшения существующих алгоритмов вычисления признаков по фрагменту звукозаписи;
2. исследовать применимость некоторых универсальных методов глубокого обучения к получению музыкальных признаков;
3. улучшить алгоритм определения аккорда по вектору признаков, использующий сопоставление с шаблонами аккордов;
4. реализовать описанные алгоритмы в виде комплекса программ, позволяющего распознавать последовательность аккордов в поданном на вход звуковом файле;
5. сравнить качество распознавания аккордов с аналогами.

Научная новизна и основные положения, выносимые на защиту:

1. Реализован новый метод распознавания последовательности аккордов в звукозаписи, не использующий алгоритмов машинного обучения.
2. Реализован новый метод представления звукозаписи в виде последовательности векторов признаков с применением многослойной нейронной сети.
3. Проведён сравнительный анализ результатов работы предлагаемых методов на коллекции из 319 звукозаписей, подтверждающий их эффективность.
4. Создан комплекс программ на языках Java и Python, реализующий описанные в данной работе методы.

Практическая значимость. Разработанный метод распознавания последовательности аккордов может применяться для анализа звукозаписей с целью их самостоятельного воспроизведения, с целью поиска схожих музыкальных композиций. Метод заведомо не подвержен опасности переобучения под конкретную музыкальную коллекцию.

Апробация работы. Основные результаты диссертационной работы докладывались на всероссийской научной конференции "Анализ Изображений, Сетей и Текстов"(Екатеринбург, 2012), на всероссийской научной конференции "Анализ Изображений, Сетей и Текстов"(Екатеринбург, 2013), на конференции молодых учёных в рамках 7-й российской летней школы по информационному поиску (Казань, 2013), на 9-й международной конференции по вычислениям в

области звука и музыки (Копенгаген, 2012), на 13-й конференции международного сообщества по музыкальному информационному поиску (Порто, 2012).

Реализованный в рамках диссертации алгоритм был выставлен на соревнования среди алгоритмов распознавания аккордов MIREX Audio Chord Estimation 2012 и MIREX Audio Chord Estimation 2013 [16], [14], [15], проводимые международной лабораторией оценки систем музыкального информационного поиска (International Music Information Retrieval Systems Evaluation Laboratory) университета Иллинойса, США.

Публикации. Результаты изложены в 5 печатных изданиях, 2 из которых изданы в журналах, рекомендованных ВАК, 3 – в тезисах докладов всероссийских и международных конференций.

Все исследования, результаты которых изложены в данной работе, получены лично соискателем в процессе научных исследований. Из совместных публикаций в диссертацию включен лишь тот материал, который непосредственно принадлежит соискателю.

Объем и структура работы. Диссертация состоит из введения, пяти глав, заключения и приложения. Полный объем диссертации **88** страниц текста с **27** рисунками и 22 таблицами. Список литературы содержит **100** наименований.

Содержание работы

Во введении обосновывается актуальность исследований, проводимых в рамках данной диссертационной работы, приводится обзор научной литературы по изучаемой проблеме, формулируется цель, ставятся задачи работы, сформулированы научная новизна и практическая значимость представляемой работы.

В первой главе приводятся теоретические сведения о звуке и его свойствах, о музыкальном строе и составе аккордов, необходимые для более конкретной постановки задач и для дальнейших построений.

Во второй главе делается обзор литературы по теме исследований. Для каждого из этапов преобразования звукозаписи в последовательность аккордов рассматриваются существующие подходы.

В третьей главе описывается реализованный в работе метод распознавания аккордов, не использующий алгоритмы машинного обучения. Метод можно условно разделить на 4 этапа, на каждом из которых реализованы новые алгоритмы.

Определение ритма позволяет более аккуратно разделить звукозапись на фрагменты. Ритм упорядочивает и группирует звуки по времени начала и продолжительности звучания. Поэтому и смена звучащего аккорда должна происходить в соответствии с ритмом. В рамках данной работы для определения ритма в зву-

козаписях в качестве вспомогательного инструмента использовались 2 внешние библиотеки: *Beatroot* [8] и *Beat tracker* [6] из набора *Queen Mary Vamp plugins*.

В звучании любого музыкального инструмента можно условно выделить 3 части: атака, стационарная часть, затухание. В процессе атаки в музыкальном инструменте устанавливаются колебания, начинают звучать основной тон и обертоны. На протяжении стационарной части звучание меняется слабо. В процессе затухания колебания прекращаются. Для ударных инструментов атака и затухание происходят существенно раньше, чем для инструментов с выраженной высотой звучания. Поскольку ритм обычно задаётся именно ударными инструментами, имеет смысл анализировать спектр в моменты времени, отстоящие на несколько десятков миллисекунд от моментов начала метрических долей. В эти моменты звучание инструментов с выраженной высотой будет наиболее ярким и полным, в то время как ударные инструменты будут находиться в процессе затухания. Наилучшее значение для величины такой задержки определяется в пятой главе.

На каждом из фрагментов, границы которых определяются с учётом ритма и задержки, вычисляется преобразование постоянного качества, которое для дискретного сигнала $x(n)$ определяется следующим образом:

$$X[n] = \frac{1}{J(n)} \sum_{j=0}^{J(n)-1} w(n,j)x(t_j)e^{-\frac{i2\pi nj}{J(n)}}, \quad n = 0, 1, \ldots, N-1$$

Здесь $J(n)$ определяет продолжительность анализируемого фрагмента звукозаписи, $w(n,j)$ – функция, отличная от нуля на некотором промежутке – оконная функция (в работе использовалась оконная функция Хэмминга). Они зависят от номера соответствующей частотной компоненты f_n. В свою очередь, f_n можно выбрать таким образом, что каждой ступени звукоряда будет соответствовать одинаковое число частотных компонент (одна или более). Пусть N_0 – количество компонент в одной октаве, а f_{min} – частота наименьшей из анализируемых компонент. Тогда частота n-й компоненты задается формулой $f_n = 2^{n/N_0} f_{min}$. Точно так же задаются частоты для ступеней звукоряда при использовании равномерно темперированного строя, поэтому параметр f_{min} напрямую связан с частотой настройки музыкальных инструментов. Отношение $\frac{f_n}{f_{n+1}-f_n} = \frac{1}{2^{1/N_0-1}} = Q$ называется коэффициентом качества. При таком выборе частот Q не зависит от k.

Учитывая указанную связь частот компонент преобразования с частотой настройки музыкальных инструментов, её определение является еще одним важным предварительным шагом. В рамках данной работы используется алгоритм, основанный на определении количества звуковой энергии, приходящейся

на узкие частотные полосы, соответствующие разным возможным значениям отклонения частоты настройки от стандартной.

Во многих работах (например, в [1], [5]) отмечалась важность сглаживания последовательности столбцов спектрограммы или векторов признаков. Сглаживание осуществляется путём применения фильтра скользящего среднего или скользящего медианного фильтра с шириной окна w к каждой строке спектрограммы. Оно позволяет избавиться от единичных выбросов в спектре, но при этом несколько размывает спектр, снижая разрешение по времени. Если каждый столбец спектра соответствует промежутку между двумя метрическими долями, такое размытие будет слишком сильным.

Чтобы преодолеть этот недостаток, увеличим разрешение спектрограммы по времени в T раз путём вставки между каждыми моментами (t_m, t_{m+1}) перед получением спектрограммы равномерно $T-1$ промежуточных значений, где T – параметр. Тогда появляется возможность использовать достаточно большой размер окна при сглаживании, не приводящий к существенному размытию спектра во времени. После сглаживания разрешение спектрограммы уменьшается в T раз путем удаления добавленных промежуточных столбцов.

На следующем этапе к спектрограмме применяется серия преобразований. Они нацелены на акцентирование компонент, которые несут важную для идентификации аккорда информацию, и на подавление остальных компонент. Наиболее важным является подавление шума и инструментов с неопределенной высотой звучания, поскольку их спектр не зависит от звучащего аккорда и сопоставим по уровню со спектром инструментов, задающих аккорд.

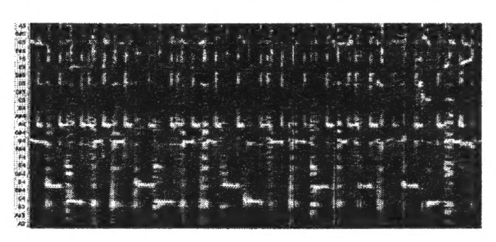


Рис. 1: Фрагмент спектрограммы *The Beatles – Love Me Do*

Как видно из рисунка 1, барабан оставляет на спектрограмме яркие вертикальные полосы. В то же время, гитаре соответствуют горизонтальные полосы. Для выделения горизонтальных линий на спектрограмме можно применить подход, используемый в обработке изображений – свёртку спектрограммы со специальным образом построенной матрицей. Будем для каждого фрагмента спектрограммы размера 9×3 с центром в точке $C_m[n]]$ вычислять его свёртку с

матрицей

$$P = \begin{pmatrix} -1 & -1 & -1 \\ -1 & -1 & -1 \\ -1 & -1 & -1 \\ 2 & 2 & 2 \\ 2 & 2 & 2 \\ 2 & 2 & 2 \\ -1 & -1 & -1 \\ -1 & -1 & -1 \\ -1 & -1 & -1 \end{pmatrix}$$

Если полученное значение больше 0, то заменим $C_m[n]$ на него, иначе – на 0. Количество строк в матрице составляет $(N_0/12) \cdot 3$. Она состоит из трёх блоков по $N_0/12$ одинаковых строк. Выше приведена матрица для значения $N_0 = 36$.

Важным свойством музыкальных звукозаписей является наличие повторений. Музыка нравится человеку в том числе из-за повторений одного и того же мотива в разных вариациях, с некоторыми изменениями. Во многих композициях имеется достаточно продолжительный повторяющийся припев. В рамках куплета может повторяться одна и та же музыкальная фраза длительностью в несколько тактов. Можно попытаться использовать повторения для улучшения спектрограммы.

В работах [19] и [5] повторяющиеся фрагменты композиции использовались для улучшения качества распознавания аккордов. В обоих методах строились матрицы самоподобия для 12-мерных хроматических векторов признаков с использованием в качестве меры подобия коэффициента корреляции Пирсона (в [19]) и евклидового расстояния (в [5]). В полученной матрице находятся линии, параллельные главной диагонали, которые соответствуют похожим друг на друга фрагментам. Эти фрагменты затем используются для дополнительного сглаживания спектрограммы.

Однако матрицу самоподобия можно строить и для столбцов спектрограммы $\{C_i\}_{i=0}^{M-1}$, каждый из которых содержит больше информации по сравнению с соответствующим вектором признаков. Обозначим эту матрицу за $\{s_{ij}\}$, где s_{ij} – евклидово расстояние между столбцами C_i и C_j. Эта матрица имеет нули на главной диагонали. Нормализуем её таким образом, чтобы $0 \leq s_{ij} \leq 1$ для всех i, j. Затем в каждой строке сохраняются $\zeta \cdot M$ наименьших значений ($0 \leq \zeta \leq 1$), а все остальные заменяются на 1. Пример полученной матрицы показан на рисунке 2.

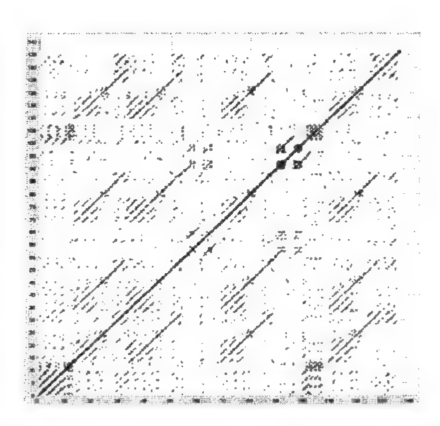


Рис. 2: Скорректированная матрица самоподобия для композиции *The Beatles – Love Me Do*

При помощи полученной матрицы можно скорректировать столбцы C_m:

$$\widehat{C}_m = \frac{\sum\limits_{j=0}^{M-1}(1 - s_{mj})C_j}{\sum\limits_{j=0}^{M-1}(1 - s_{mj})}$$

В скорректированной таким образом спектрограмме лучше выделяются границы между аккордами, что позволяет повысить качество распознавания.

На заключительном этапе необходимо классифицировать каждый из столбцов спектрограммы. Будем рассматривать в качестве множества возможных названий аккордов Y набор из названий 12 мажорных аккордов, 12 минорных и символа «N», означающего отсутствие аккорда. За основу возьмём метод ближайшего соседа с шаблонами. Эти шаблоны задаются для 12-мерных хроматических векторов. Столбцы спектрограммы охватывают несколько октав, поэтому они преобразуются в 12-мерные векторы путем суммирования значений компонент, частоты которых соответствуют ступеням звукоряда, имеющим одинаковые названия, но находящимся в разных октавах.

Для определения отсутствия звучащего аккорда применяется отдельная процедура. Участки, на которых отсутствует звучащий аккорд, не обязательно являются тишиной. Это может быть также соло на ударных инструментах или какой-либо шум. Поэтому нельзя определять отсутствие аккорда как близость столбца спектрограммы или вектора признаков к нулевому вектору. Необходимо каким-то образом учитывать информацию о наличии и выраженности звуков, соответствующих ступеням звукоряда.

Вычислим для каждого столбца спектрограммы следующие величины:

1. отношение максимальной по значению из компонент к сумме значений всех компонент;
2. отношение суммы значений компонент, частоты которых в точности соответствуют частотам ступеней звукоряда, к сумме значений всех компонент.

После чего будем определять отсутствие звучащего аккорда на фрагменте звукозаписи, соответствующем данному столбцу спектрограммы, если произведение указанных двух величин будет меньше некоторого порогового значения L_N. Кроме того, будем считать, что аккорды отсутствуют на участках до первой и после последней из определённых ранее метрических долей.

В результате экспериментов было обнаружено, что некоторые последовательности аккордов являются маловероятными в реальной композиции, и скорее всего является ошибочными. Для двух классов таких последовательностей предлагается метод их исправления.

К первому классу относятся последовательности, в которых аккорды имеют общую основную ноту, но различные типы, например: A:maj-A:min-A:maj-A:min. Появление таких последовательностей возможно, поскольку соответствующие векторы признаков достаточно близки друг к другу. Для каждой такой последовательности находится вектор признаков, являющийся средним арифметическим составляющих её векторов. Аккорд, соответствующий полученному вектору признаков, приписывается всей последовательности.

Ко второму классу относятся последовательности из 3 разных идущих подряд аккордов. Допустим, в результате работы алгоритма была получена последовательность аккордов A-B-C (при этом возможно A=C). В этом случае более вероятно, что на самом деле имел место один из следующих 4 вариантов: A-A-C, A-C-C, A-B-B, B-B-C. Из них выбирается тот, для которого сумма расстояний от векторов признаков до соответствующих шаблонных векторов минимальна. Очевидно, что такая коррекция будет ошибочной в тех случаях, когда аккорд действительно звучит только в течение одной метрической доли.

В четвёртой главе описывается метод получения тональных признаков при помощи многослойной нейронной сети. При таком подходе не используется большая часть из описанных в третьей главе преобразований спектра. Вместо этого каждый из столбцов спектрограммы подаётся на вход нейронной сети, на выходах которой получается вектор признаков.

В 2012 году Хамфри в [11] предложил использовать свёрточные нейронные сети для получения признаков, позволяющих классифицировать звучащий аккорд. В данной работе рассматриваются обычные многослойные (в том числе рекуррентные) нейронные сети, предварительно обучаемые с помощью очищающих автоассоциаторов. Такие сети были успешно использованы для распозна-

вания речи в [18]. Похожий подход был продемонстрирован в работе [3], где рекуррентная нейронная сеть возвращает на выходе сразу распознанный аккорд, который при помощи рекуррентного соединения подаётся на вход на следующем шаге.

Даётся определение автоассоциатора (автоэнкодера) как пары нелинейных преобразований, первое из которых переводит исходный вектор в сжатое или разреженное внутреннее представление, а второе из этого представления восстанавливает исходный вектор. Эту пару преобразований удобно представлять в виде нейронной сети с одним внутренним слоем. В процессе обучения её параметры настраиваются таким образом, что во внутреннем слое получается эффективное сжатое или разреженное представление входного вектора. Это представление оказывается ещё более эффективным, если на вход подаётся искажённое, зашумлённое представление исходного вектора [23]. Соответствующим образом модифицированный автоассоциатор получил название очищающего.

Из автоассоциаторов можно строить многослойные модели, отождествляя нейроны из скрытого слоя одного автоассоциатора со входными нейронами другого. В полученной модели слои можно обучать друг за другом на неразмеченных данных. Значения, полученные в скрытом слое последнего из автоассоциаторов, могут быть использованы как векторы признаков. Пример полученной нейронной сети показан на рисунке 3

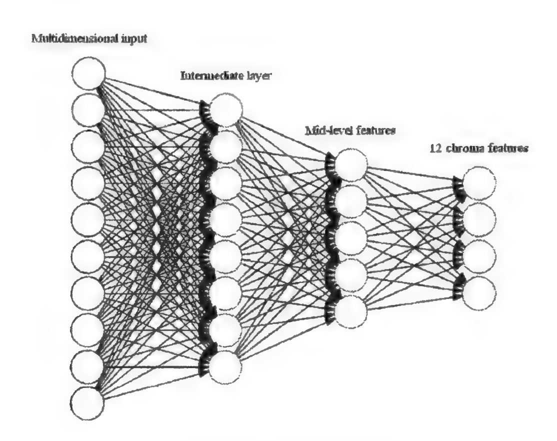


Рис. 3: Многослойная нейронная сеть

Существенным недостатком автоассоциаторов является невозможность содержательной интерпретации значений во внутреннем слое. В частности, невозможно построить шаблонные наборы значений, соответствующие аккордам. Можно попытаться обучить алгоритм классификации на векторах значений

на выходах внутреннего слоя. Но для случая 25 классов и достаточно большой размерности векторов для обучения такого классификатора может потребоваться слишком много данных.

Вместо этого соединим внутренний слой с дополнительным слоем, имеющим 12 выходов. Полученную нейронную сеть обучим на размеченных данных таким образом, чтобы на выходе получались хроматические векторы. На вход этой сети будут подаваться столбцы спектрограммы. Таким образом, вместо задачи классификации полученная нейронная сеть решает задачу регрессии. Классификация же полученных 12-мерных векторов делается точно так же, как в третьей главе.

Предварительное обучение слоёв-автоассоциаторов будем производить методом мини-пакетного (mini-batch) стохастического градиентного спуска. При этом сначала обучается первый слой на всём обучающем множестве, затем полученные значения используются для обучения второго слоя, и так далее. Окончательное обучение сети в целом также производится методом мини-пакетного стохастического градиентного спуска. При этом в качестве целевых векторов используются 12-мерные бинарные шаблоны аккордов, а для случая отсутствия аккорда – нулевой вектор. При классификации, соответственно, отсутствие аккорда определяется в случае, когда ни одна из компонент полученного вектора по абсолютной величине не превосходит некоторого значения Δ, которое подбирается опытным путём.

В случае, когда в обучающей выборке большинство примеров соответствуют только одному классу, очень вероятно получить в итоге сеть, которая все векторы будет классифицировать как принадлежащие этому классу. В данном случае имеется 25 возможных классов, и желательно иметь приблизительно одинаковое количество примеров на каждый класс. Однако не все аккорды используются в музыке одинаково часто, и, например, аккорд *до-мажор* может встречаться в обучающей выборке в разы чаще, чем *фа-диез-мажор*.

Аналогичная проблема встречается и при обучении скрытых марковских моделей и байесовских сетей для определения последовательности аккордов по последовательности векторов признаков. В этих моделях часто используют циклический сдвиг векторов признаков для усреднения параметров, соответствующих разным аккордам. Этот процесс подробно описан, например, в [24]. Идея его состоит в том, что, поскольку в хроматическом векторе каждая компонента соответствует одному тональному классу, его циклический сдвиг даёт вектор, соответствующий аккорду того же типа (мажорный или минорный) с основной нотой, сдвинутой на полутон.

В данном случае при окончательном обучении нейронной сети в целом можно также использовать сдвиг. Но входными векторами являются столбцы

спектрограммы, и циклический сдвиг соответствует неестественному переносу высокочастотных компонент в область низких частот (или наоборот). Циклический сдвиг можно эмулировать, добавив одну октаву к частотному диапазону спектрограммы, после чего просто сдвигать по столбцу спектрограммы окно размером на октаву меньше.

При помощи такого сдвига из каждого столбца спектрограммы получается 12 различных столбцов, соответствующих 12 аккордам одного типа с разными основными нотами. Это позволяет уравновесить количество аккордов в пределах одного типа. Чтобы уравновесить количество аккордов между типами, потребуем, чтобы в процессе генерации обучающей выборки из спектрограмм разница между общим количеством мажорных аккордов и общим количеством минорных аккордов не превосходила некоторого заданного числа.

Во время тестирования также можно использовать циклический сдвиг. Для каждого столбца спектрограммы генерируется 12 тестовых векторов, каждый из которых при помощи нейронной сети преобразуется в хроматический вектор. Для полученных 12 хроматических векторов производятся соответствующие обратные сдвиги, а в качестве результата берется среднее арифметическое от полученных векторов.

Пятая глава посвящена экспериментам, нацеленным на подбор оптимальных параметров метода и определение степени влияния каждого из шагов метода. Описываются используемые метрики качества распознавания аккордов, в соответствии с которыми возможно сравнивать разные методы или разные варианты одного метода. Для проверки наличия статистически значимой разницы между разными вариантами метода предлагается использовать непараметрический критерий Фридмана. Приводятся некоторые количественные характеристики тестовой коллекции. Формулируются основные типы ошибок, позволяющие более точно определить влияние различных параметров метода на качество распознавания аккордов.

Далее поочерёдно рассматривается эффект от выбора различных значений параметров и различных алгоритмов на промежуточных этапах. Вычисляются значения метрик на всей коллекции, приводятся диаграммы ошибок.

По результатам экспериментов сделаны следующие выводы.

1. Предложенные и реализованные в работе повышение разрешения спектрограммы по времени и задержка относительно моментов начала метрических долей существенно улучшают качество распознавания аккордов.

2. Эффект от определения ритма существенно выше, чем от определения частоты настройки.

3. Использование свёртки для очистки спектрограммы не улучшает качество распознавания аккордов.

4. Применение информации о самоподобии позволяет существенно улучшить результат.

5. Применение нейронной сети для получения хроматических признаков позволяет добиться результата чуть более слабого, чем для описанного в третьей главе алгоритма. При этом наличие рекуррентных соединений практически не влияет на качество распознавания аккордов.

6. Из реализованных эвристик полезной оказывается только одна, связанная с исправлением аккордов длительностью в одну метрическую долю.

7. Реализованный в рамках работы метод позволяет добиться качества распознавания, сравнимого с наилучшими существующими в мире аналогами. При этом время обработки звукозаписей у реализованного метода остаётся небольшим даже с использованием обычного компьютера. Быстродействие метода можно повысить при незначительном снижении качества распознавания аккордов.

8. При использовании нейронной сети для получения признаков совокупная продолжительность работы метода существенно увеличивается даже при наличии только одного скрытого слоя.

9. Описанный в главе 3 алгоритм позволяет правильно распознать аккорды в среднем для 77.5% от продолжительности композиции при использовании только мажорных и минорных трезвучий (для используемого тестового набора из 318 композиций групп *The Beatles*, *Queen*, *Zweieck*).

В **заключении** приведены основные результаты работы, которые заключаются в следующем:

1. На основе анализа свойств музыкальных звукозаписей был разработан метод для более точного выделения в звуке компонент, соответствующих музыкальным инструментам.

2. Исследование показало, что глубокие нейронные сети в применении к получению музыкальных признаков могут показывать хорошие результаты, сравнимые с результатами традиционных подходов к получению признаков.

3. Численные исследования показали, что реализованные в рамках работы подходы позволяют добиться качества распознавания аккордов, сравнимого

с лучшими из известных алгоритмов, при достаточно высокой скорости обработки.

4. Для выполнения поставленных задач был создан программный комплекс на языках Java и Python, свободно доступный через интернет.

Список цитируемой литературы

1. Analyzing Chroma Feature Types for Automated Chord Recognition / Nanzhu Jiang, Peter Grosche, Verena Konz, Meinard Müller // Proceedings of the AES 42nd International Conference: Semantic Audio. — Ilmenau, Germany : AES, 2011. — P. 285–294.

2. Aono Y., Katayose H., Inokuchi S. A Real-time Session Composer with Acoustic Polyphonic Instruments // Proceedings of ICMC 1998. — 1998. — P. 236–239.

3. Boulanger-Lewandowski Nicolas, Bengio Yoshua, Vincent Pascal. Audio chord recognition with recurrent neural networks // Proceedings of the 14th International Society for Music Information Retrieval Conference. — 2013. — November 4-8. — http://www.ppgia.pucpr.br/ismir2013/wp-content/uploads/2013/09/243_Paper.pdf.

4. Brown Judith, Puckette Miller S. An efficient algorithm for the calculation of a constant Q transform // Journal of the Acoustical Society of America. — 1992. — November. — Vol. 92, no. 5. — P. 2698–2701.

5. Cho Taemin, Bello Juan P. A Feature Smoothing Method for Chord Recognition Using Recurrence Plots // Proceedings of the 12th International Society for Music Information Retrieval Conference. — Miami (Florida), USA, 2011. — October 24-28. — P. 651–656. — http://ismir2011.ismir.net/papers/OS8-4.pdf.

6. Davies Matthew E. P., Plumbley Mark D. Context-Dependent Beat Tracking of Musical Audio // Trans. Audio, Speech and Lang. Proc. — 2007. — March. — Vol. 15, no. 3. — P. 1009–1020. — URL: http://dx.doi.org/10.1109/TASL.2006.885257.

7. De Haas W. Bas, Magalhães José Pedro, Wiering Frans. Improving Audio Chord Transcription by Exploiting Harmonic and Metric Knowledge // Proceedings of the 13th International Society for Music Information Retrieval Conference. — Porto, Portugal, 2012. — October 8-12. — http://ismir2012.ismir.net/event/papers/295-ismir-2012.pdf.

8. Dixon Simon. Evaluation of the Audio Beat Tracking System BeatRoot // Journal of New Music Research. — 2007. — March. — Vol. 36, no. 1. — P. 39–50. — URL: http://dx.doi.org/10.1080/09298210701653310.

9. Fujishima Takuya. Realtime Chord Recognition of Musical Sound: a System Using Common Lisp Music // Proc. ICMC, 1999. — 1999. — P. 464–467. — URL: http://ci.nii.ac.jp/naid/10013545881/en/.

10. Gómez E. Tonal Description of Music Audio Signals : Ph. D. thesis / E. Gómez ; Universitat Pompeu Fabra. — 2006. — URL: files/publications/emilia-PhD-2006.pdf.

11. Humphrey E.J., Cho T., Bello J.P. Learning a robust tonnetz-space transform for automatic chord recognition // Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP-12). — Kyoto, Japan, 2012. — May. — P. 453–456.

12. Khadkevich Maksim, Omologo Maurizio. Time-frequency reassigned features for automatic chord recognition // Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP 2011, May 22-27, 2011, Prague Congress Center, Prague, Czech Republic. — IEEE, 2011. — P. 181–184.

13. Lee K. Automatic chord recognition from audio using enhanced pitch class profile // ICMC Proceedings. — 2006.

14. MIREX 2013: Audio Chord Estimation Results Billboard 2012. — 2014. — March. — URL: http://www.music-ir.org/mirex/wiki/2013:Audio_Chord_Estimation_Results_Billboard_2012.

15. MIREX 2013: Audio Chord Estimation Results Billboard 2013. — 2014. — March. — URL: http://www.music-ir.org/mirex/wiki/2013:Audio_Chord_Estimation_Results_Billboard_2013.

16. MIREX 2013: Audio Chord Estimation Results MIREX 2009. — 2014. — March. — URL: http://www.music-ir.org/mirex/wiki/2013:Audio_Chord_Estimation_Results_MIREX_2009.

17. MIREX Home Page. — 2014. — March. — URL: http://www.music-ir.org/mirex/wiki/MIREX_HOME/.

18. Maas A. Le Q. O'Neil T. Vinyals O. Nguyen P. Ng A. Recurrent Neural Networks for Noise Reduction in Robust ASR // Proceedings of INTERSPEECH (2012). — 2012.

19. Mauch Matthias, Dixon Simon. Approximate Note Transcription for the Improved Identification of Difficult Chords // Proceedings of the 11th International Society for Music Information Retrieval Conference. — Utrecht, The Netherlands, 2010. — August 9-13. — P. 135–140. — http://ismir2010.ismir.net/proceedings/ismir2010-25.pdf.

20. Mauch Matthias, Noland Katy, Dixon Simon. Using Musical Structure to Enhance Automatic Chord Transcription // Proceedings of the 10th International Society for Music Information Retrieval Conference. — Kobe, Japan, 2009. — October 26-30. — P. 231–236. — http://ismir2009.ismir.net/proceedings/PS2-7.pdf.

21. Müller Meinard, Ewert Sebastian, Kreuzer Sebastian. Making chroma features more robust to timbre changes // Proceedings of the 2009 IEEE International Conference on Acoustics, Speech and Signal Processing. — ICASSP '09. — Washington, DC, USA : IEEE Computer Society, 2009. — P. 1877–1880. — URL: http://dx.doi.org/10.1109/ICASSP.2009.4959974.

22. Oudre Laurent, Grenier Yves, Févotte Cédric. Template-Based Chord Recognition : Influence of the Chord Types // Proceedings of the 10th International Society for Music Information Retrieval Conference. — Kobe, Japan, 2009. — October 26-30. — P. 153–158. — http://ismir2009.ismir.net/proceedings/PS1-17.pdf.

23. P. Vincent H. Larochelle I. Lajoie Y. Bengio, Manzagol P.-A. Stacked Denoising Autoencoders: Learning Useful Representations in a Deep Network with a Local Denoising Criterion // The Journal of Machine Learning Research. — 2010. — Vol. 11. — P. 3371–3408.

24. Sheh Alexander, Ellis Daniel P. W. Chord segmentation and recognition using EM-trained hidden markov models. // ISMIR. — 2003.

Публикации в научных изданиях по теме диссертации

Статьи в рецензируемых научных журналах, определенных ВАК.

25. Н.Ю. Глазырин. О задаче распознавания аккордов в цифровых звукозаписях // Известия Иркутского государственного университета, серия «Математика». — 2013. — Т. 6, № 2. — С. 2–17.

26. Glazyrin Nikolay. Mid-level Features for Audio Chord Estimation using Stacked Denoising Autoencoders // Ученые записки Казанского университета. Серия Физико-математические науки. — 2013. — Vol. 155, no. 4. — P. 109–117.

Свидетельства о государственной регистрации программ для ЭВМ.

27. Глазырин Н. Ю. Chordest. — Свидетельство № 2014614132 от 17.04.2014. — 2014.

Другие публикации

28. Глазырин Н.Ю. Клепинин А.В. Выделение гармонической информации из музыкальных аудиозаписей // Доклады всероссийской научной конференции «Анализ Изображений, Сетей и Текстов» (АИСТ 2012). — Екатеринбург, Россия, 2012. — С. 159–168.

29. Глазырин Н.Ю. Клепинин А.В. Применение автоассоциаторов к распознаванию последовательностей аккордов в цифровых звукозаписях // Доклады всероссийской научной конференции «Анализ Изображений, Сетей и Текстов» (АИСТ 2013). — Екатеринбург, Россия, 2013. — С. 199–203.

30. Glazyrin Nikolay, Klepinin Alexander. Chord Recognition using Prewitt Filter and Self-Similarity // Proceedings of the 9th Sound and Music Computing Conference. — Copenhagen, Denmark, 2012. — July 11-14. — P. 480–485.

Подписано в печать 10.12.2014 Формат 60х84 1/16
Бумага офсетная. Усл. печ. л. 0,92
Тираж 50 экз. Заказ № 1741

Отпечатано в типографии ИПЦ УрФУ
620000, Екатеринбург, ул. Тургенева,4